基于 VSM 的美国一流大学图书馆网站导航文本调查与分析

尹相权　李书宁
(北京师范大学图书馆　北京　100875)

摘要: 【目的】通过分析美国一流大学图书馆网站导航文本的特点，为国内大学图书馆导航建设提供建议。【方法】结合一流大学应具有一流学科、一流教师和一流学生的思路选取 15 所美国一流大学图书馆，基于标签云和文本挖掘模型 VSM，分析导航文本词维度上的共性和特异性，并结合《2016 年美国图书馆状况报告》进行数据验证。【结果】与人工调研相比，基于 VSM 模型的统计方法可以更直观、快速地给出基础调研结果，调研结果可供进一步深度文本分析参考。【局限】仅选取一级导航、二级导航和首页标题板块概况导航文本。【结论】基于文本数据挖掘模型的统计方法可以更直观、快速地给出基础调研结果，为高校图书馆网站导航建设提供参考。

关键词: 美国大学图书馆　网站导航　向量空间模型　调查分析
分类号: G250.7

1　引　言

互联网时代，图书馆网站转变为图书馆内容展示和服务的主要窗口，通过图书馆网站获取信息已成为用户利用图书馆资源的一种基本方式。随着近年图书馆信息资源量和服务内容的增加，包括提供文献资源、提供图书馆服务介绍信息、提供图书馆消息、促进与读者的交流互动等，图书馆网站要承载的内容远远超出了原有的限度。而用户的需求也日趋多样化，要能够在图书馆网站以最快的速度找到自己所需要的东西，由于图书馆网站所承载的内容有限，且往往有很多相对独立的资源管理系统与服务模块，这势必会造成相当多的内容被隐藏，图书馆服务无法更好地为读者所用，读者也无法知道图书馆还有哪些服务。在这种情况下，作为提供给用户的最直接、最方便的网站内容访问工具，图书馆网站导航就变得非常有意义，而且伴随着信息超载，这种导航将越来越有意义。图书馆网站导航主要指位于网页页眉区域的，在页眉横幅图片上边或下边的一排水平导航文字，它起着链接图书馆网站的各个页面的作用。另外，考虑到首页亦可提供一个简便快捷的操作入口，帮助用户快速定位到所需要的资源，图书馆首页各个栏目的标题也应该纳入图书馆导航的范畴。

图书馆网站应完善导航功能，有针对性地把物理上分散的、杂乱无章的信息资源重新组织，使网络用户能够快捷地找到自己所需要的信息[1]。国内著名大学在建设图书馆网站时，大多参考了国外著名大学的图书馆网站。以清华大学图书馆为例，其在改版时将图书馆网站的信息和资源的组织、揭示与布局作为重要课题，并参考了国外著名大学图书馆网站[2-3]。但是国内著名大学图书馆大多重点描述了参考国外大学图书馆网站之后的实施，没有对导航调研方法进行描述和分析。因此，本文选取 15 所美国一流大学图书馆的导航文本，借助向量空间模型 VSM 进行研究，以期为国内图书馆网站导航建设提供参考。

2　美国一流大学图书馆导航文本分析

2.1　数据收集

按照一流大学应具有一流学科、一流教师和一流

通讯作者：尹相权，ORCID：0000-0002-9815-896X，E-mail：yinxq@lib.bnu.edu.cn。

学生的思路选取 10 所美国一流大学[4]。另外，考虑到美国大学图书馆显著的社会服务特点，为考察图书馆的社会服务，在 10 所大学的基础上，又增加了 5 所大学[5]。最终选取的 15 所大学图书馆分别为：哈佛大学图书馆、斯坦福大学图书馆、麻省理工学院图书馆、耶鲁大学图书馆、普林斯顿大学图书馆、哥伦比亚大学图书馆、芝加哥大学图书馆、加州理工学院图书馆、宾夕法尼亚大学图书馆、加州大学伯克利分校图书馆、康奈尔大学图书馆、加州大学戴维斯图书馆、田纳西大学图书馆、北卡罗莱纳州立大学图书馆以及西北大学图书馆。

在 Peter Morville 提出的用户体验蜂巢模型(User Experience Honeycomb)中，可寻性是用户体验的主要指标之一，主要通过导航与定位体现[6]。依据其理论，一级导航、二级导航和首页板块标题可以代表一个网站内容的主要导航组织方式，因此，本文选取一级导航、二级导航和首页板块标题文本作为分析对象。

首先，人工收集这 15 所大学图书馆的一级导航、二级导航和首页板块标题，并进一步根据导航文字的语义进行归一化处理，即清洗和转换工作。例如，“About”、“ABOUT”、“About us”和“About the library”，经统一大小写和文本替换工作后，统一文字为“about us”; “Help”和“Get Help”统一为“get help”; 拆分并列词组，把“Search & Find”拆分为两个词条“search”和“find”，把“tools for prospective students| current students|faculty or staff|alumni or friends” 拆分为“tools for prospective students”、“tools for current students”、“tools for faculty or staff”和“tools for alumni or friends”。针对不同层次的分析目标，提供一级导航文字统计分析、首页板块统计分析和导航文字在导航词维度上的统计分析。

2.2　文本统计分析方法

在对文本进行统计分析时，主要考察文本的特性以及文本与文本之间的相似性。为了更清晰地对导航文本进行解析，本文首先利用标签云对导航文本整体进行直观描绘，之后，通过统计方法对文本进行特征抽取，用于表征各个文本的特性，并进一步在文本表征的基础上进行文本相似度计算，挖掘文本和文本之间隐含的相似性。每个文本的特征向量可通过降维处理并打印，形成文本的摘要信息，供特性分析使用。

标签云是一套相关的标签以及与此相应的权重。权重影响使用的字体大小或其他视觉效果。标签字体越大，此条目在网站中出现的次数越多。标签云在直观展示网站的显著内容时十分适用，可用于各个网站导航之间的共性分析。通常典型的标签云有 30 至 150 个标签，当文本较多时，标签云的直观性会降低。

本文采用的文本相似性计算方法为较为经典的向量空间模型(Vector Space Model, VSM)。VSM 将对文本内容的处理简化为向量空间中的向量运算，并且它以空间上的相似度表达语义的相似度[7]。其中，导航文本的相似度通过两个多维向量的夹角的余弦值来表征。两个向量的夹角越小，余弦值越高，代表导航文本之间相似度越高。

具体而言，假设有 M 个导航文本，对每个导航文本进行特征提取，假设特征为 N 维，可以得到 M×N 的特征向量矩阵 F，将其映射到 VSM 模型中，可以得出 M 个文本中任意两个文本的特征距离。本文采用向量的余弦距离作为它们之间的语义距离。假设有导航文本 A 和导航文本 B，则它们的语义距离如公式(1)所示。

$$dis(A,B)=\cos\left(\frac{\overline{A}\cdot\overline{B}}{|\overline{A}|\cdot|\overline{B}|}\right)=\cos\left(\frac{\sum_{i=1}^{N}a_ib_i}{\sqrt{\sum_{i=1}^{N}a_i^2\cdot\sum_{i=1}^{N}b_i^2}}\right) \quad (1)$$

本文选择 TF-IDF(Term Frequency–Inverse Document Frequency)方法。TF-IDF 的主要思想是：如果某个词或短语在一个文本中出现的频率高，并且在其他文本中很少出现，则认为此词或者短语具有很好的类别区分能力。

VSM 处理的文本需要分词、去停用词等操作，考虑到导航文字中词条和词条之间有明显物理间隔，本文把一组导航词作为一个词组进行处理，首先对词组进行归一化预处理，并计算每个词组的 TF-IDF 值，由此，各个网站内容的表示形式转换为词组的 TF-IDF 值向量，基于该向量进行相似度计算并选取相似度阈值，挖掘各个导航文本之间潜在的关联，并选取各导航文本 TF-IDF 值最高的 10 维特征，打印为导航文本的关键词组，分析各个网站导航的特性。

2.3　结果分析

首先利用标签云分别直观展示一级导航文字共性

和首页板块标题的共性。之后，为进一步探索具体导航文本之间的相似性，利用 VSM 对每个网站的导航文字，包括一级导航文字、二级导航文字、首页板块标题文字，进行语义相似度分析，并打印出各个大学图书馆导航文本的特征向量，结合《2016 年美国图书馆状况报告》[8]进行分析。

(1) 导航直观共性分析

在对导航词进行预处理之后，对 15 所一流大学图书馆的一级导航文字进行汇总，导入开源标签云生成器 TAGUL[9]中，生成一级导航标签云，如图 1 所示。可见，美国一流大学网站中，关于我们(about us)、研究支持(research support)、服务(services)、帮助(get help)、数据集(collections)、图书馆们(libraries)、研究(research)、检索(search)和寻找(find)占据了显著位置，说明这些导航词在各大学图书馆一级导航中出现的频率较高。



图 1　美国 15 所一流大学图书馆一级导航标签云

同样，利用首页板块标题文字生成首页板块标题标签云，如图 2 所示，新闻(news)、检索(search)、发现(find)、活动(events)模块是比较通用的模块。



图 2　美国 15 所一流大学图书馆首页板块标题标签云

(2) 基于 VSM 的导航共性分析

基于 VSM 的网站余弦相似度计算结果表明，各个网站之间的相似度偏低(均低于 0.30)。说明各个大学图书馆在建设网站导航时，并不存在过度借鉴现象，这与实际情况相符合，也说明了本文选取的文本相似度计算方法的可靠性。同时，如导航文本直观共性分析所示，有些导航词会在多个大学图书馆导航中出现，为了进一步探索网站之间的共性，选定相似度阈值 0.20，发现 7 对组合存在弱相似性，为了直观表示，用实线关联具有一定相似性的高校，如图 3 所示，哈佛大学图书馆与斯坦福大学图书馆、麻省理工学院图书馆等具有一定的弱相似性，从打印的特征表示向量中可以看出，图书馆们(libraries)、活动(events)和员工目录(staff directory)等导航词对相似度值有一定贡献。

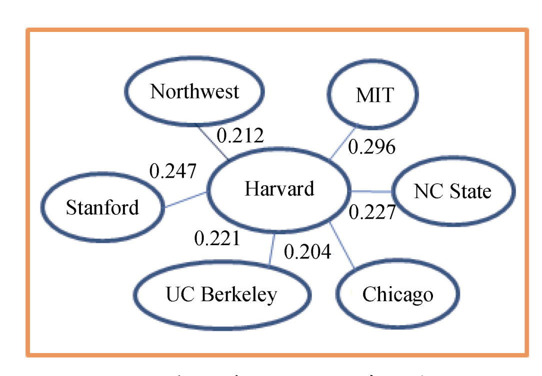


图 3　与哈佛大学图书馆具有弱相关性的 6 所大学图书馆及相似度

(3) 基于 VSM 的导航特性分析

为了进一步考察各个大学导航文本的特性，本文打印了各个大学的导航词组特征向量，按照词的 TF-IDF 值从高到低的顺序，选取排序前 10 位的特征词组进行人工分析，如表 1 所示。

共性(将超过 1/3 大学存在的特征词作为共性)：在 15 所大学中，有 11 所大学把图书馆们(libraries)作为重点展示对象，其中，有 10 所大学的第一位特征词均为图书馆；7 所大学中检索(search)相关特征词显著；6 所大学有研究指南特征词(guides)；6 所大学有活动通知(events)，其中有 2 处显示为 news and events；6 所大学显著提供了员工指南(staff directory)；5 所大学有关于我们(about us)特征词。通过这些共性导航词，不难发现，一流大学图书馆大多把展示资源(图书馆、员工、关于我们)和服务(检索、研究指南)放在显著位置。与

表 1　15 所一流大学图书馆 Top10 导航文字

图书馆	Top10 导航文字
斯坦福大学	libraries, access for persons with limited mobility, jobs, collecting areas, events, computing(equipment & services), privileges, search tools, chat, course guides
康奈尔大学	faq for instructors, research data management services, equipment, computing, how to submit course reserves, library spaces, help, search tips: catalog, research guides, search
耶鲁大学	libraries, guide to using special collections, get it @yale (borrow direct, interlibrary loan, scan & deliver), find ejournals by title, search worldcat, policies, elischolar, search, search library catalog (orbis), services for persons with disabilities
哥伦比亚大学	recommend a title for purchase, deposit your research, technology, borrow direct, interlibrary loan, computing, policies, butler library lockers, room reservation, study spaces
加州大学伯克利分校	libraries, hours and maps, reserve a study room, news and events, renew, staff directory, research help, how to find, online exhibits, about us
芝加哥大学	libraries, copyright info, borrowdirect, employment, database finder, chapters, privileges, other local collections, research centers, library surveys
北卡罗莱纳州立大学	libraries, google scholar, interlibrary loan, staff directory, filmfinder, tours, search, give to the library, course reserves, visitor information
西北大学	libraries, resources, contact the library, tools for alumni, events, interlibrary loan, search tools, visit, guides, tools for graduate students
普林斯顿大学	libraries, today’s hours, borrow direct, study spaces and lockers, news and events, recommend a purchase, staff directory, new catalog, about us, research guides
加州理工学院	publish on demand, site map, archives, how do i...?, caltech open access policy faq, software available, about us, ask a librarian, friends of the caltech libraries, copyright support
哈佛大学	libraries, initiatives, archives, contribute your research, events, e-resources, resources for alumni, departments, staff directory, get it services
田纳西大学	ut dissertations, employment, give to the libraries, the library society, staff directory, libraries a-z, renew items, citing sources, music library, research guides
宾夕法尼亚大学	libraries, resources, penn’s libraries, create a video, staff directory, subjects/collections, tools, tutorials for tools, search, digitalpenn
麻省理工学院	scholarly publishing, galleries, tip faq, events, use policy, more search options, your account, about us, citation software, study spaces
加州大学戴维斯分校	libraries, melvyl, request a book/article, subject guides, digital scholarship, engineering, borrowing/circulation, carlson health sciences, about us, exhibits

国内大学不同的是，这 15 所一流大学图书馆大多把活动通知与新闻分开展示，并且把活动通知放在更显著的位置。

通过观察表 1 中的各网站导航的前 10 维特征向量，本文总结了其体现的部分特色服务，如下:

①斯坦福大学图书馆的特征向量中，位列第 2 的特征词(组)为“access for persons with limited mobility”，展示了该网站对残疾人的支持服务；位列第 6 的特征词：computing (equipment & services)展示了其对计算相关服务的支持。

②耶鲁大学图书馆，与斯坦福大学类似，对残疾人的服务也有所体现(services for persons with disabilities)。另外，耶鲁大学图书馆的特色——Orbis 在线图书馆目录系统(Online Library Catalog)也出现在特征向量中。

③哈佛大学图书馆和西北大学图书馆分别有显著的为校友提供的信息，其中，哈佛大学体现在第 8 位特征词，西北大学为第 4 位。

④加州大学戴维斯分校图书馆(UC Davis)的在线图书馆目录系统 Melvyl 体现在了第 2 位。另外，第 8 位特征词(carlson health sciences)体现其医疗健康相关特色数据。

⑤麻省理工学院图书馆的学术出版服务(scholarly publishing)位列特征向量的第 1 位。

这些特征向量表明，各大学图书馆均有其服务侧重点，可满足不同用户、不同领域的技术需求，例如斯

坦福大学的计算相关服务、耶鲁大学的 Orbis 在线图书馆目录系统、普林斯顿大学的新目录系统、加州大学戴维斯分校的 Melvyl 在线图书馆目录系统、加州大学戴维斯分校的医疗健康相关数据、麻省理工学院的学术出版服务等。另外，各个大学图书馆在开展社会服务方面各有特色，包括残疾人服务和校友服务等。以上结果也印证了《2016 年美国图书馆状况报告》中的表述：图书馆在积极开展服务转型以满足用户技术需求[8]。

此外，上述报告中也提到，调查表明学生和教职工认可大学图书馆在展示研究技术、增强学生读写能力和管理课程资源等方面的价值。大学图书馆正在通过科技界和数字化学术中心探寻激励学生成功的创新方法[8]。以上内容在本文的数据中也能找到对应的数据支持。研究技术(research techniques)：所有大学图书馆均有涉及，具体体现为帮助、指南、目录、检索和发现等；管理课程资源(course reserves)：康奈尔大学(第 5 位)、北卡罗莱纳州立大学(第 9 位)、斯坦福大学(第 10 位)；科技界(publishing on demand)：加州理工学院(第 1 位)、麻省理工学院(第 1 位)；数字化学术中心(digital scholarship)：加州大学戴维斯分校(第 5 位)。

总之，通过特征向量，可以迅速了解各大学图书馆导航文本之间的共性和特异性，并进一步为解析图书馆当前重点以及发展方向提供数据基础。需要注意的是，特征向量表示的是网站导航的代表性文本而非全部文本。例如，完整的网站导航文本中，除了斯坦福大学和耶鲁大学、北卡罗莱纳州立大学、加州大学伯克利分校、哥伦比亚大学、康奈尔大学等均有提及残疾人相关服务，但是只有前者体现在 Top10 导航文字中。

2.4 对国内高校图书馆网站导航建设的启示

调研结果对国内高校图书馆网站导航建设有如下启示：

(1) 各个大学图书馆在建设网站导航时，可应用但不仅限于以下共性导航词：一级导航词可以考虑选取关于我们、研究支持、服务、帮助、数据集、图书馆们、研究、检索和寻找等共性导航词；在首页板块中，可以选取新闻、检索、发现和活动等共性板块；在整个导航文本中，图书馆们、检索、研究指南、活动、员工目录和关于我们具有一定的普遍性，可供选取。

(2) 在建设导航过程中，应避免过度借鉴现象，一方面在导航词中加入图书馆特色元素，例如，如果图书馆有特色编目，可在编目导航词中加入特色编目简称。另一方面可以参考国外大学，在导航中区分用户类别，例如残疾人、校友、学生等。更重要的是，应顺应时代发展的要求，重点拓展特色服务，包括社会服务，激励学生创新的技术服务等。

3 结 语

本文以 15 所美国一流大学图书馆的网站导航文字作为调研对象，基于 VSM 模型对导航文本进行导航词维度上的统计，直观展示了各网站导航的共性和特异性，并结合《2016 年美国图书馆状况报告》进行数据验证。通过分析发现：

一级导航文本中，多数图书馆重点在于展示其资源和服务，包括关于我们、研究支持、服务、帮助、数据集、图书馆们、研究、检索和寻找等；在首页板块中，新闻和事件、检索和快速链接模块是比较通用的模块。

网站整体导航文本之间的相似度偏低，各个网站的特性向量有显著差异，在给定相似度阈值的条件下，只有哈佛大学图书馆与斯坦福大学图书馆、麻省理工学院图书馆等 6 所大学图书馆具有一定的弱相似性；图书馆们、检索、研究指南、活动、员工目录和关于我们等特征词具有一定的普遍性；从各个特征向量中，可以找到以残疾人服务为例的相关社会服务特征词，并有特征词与《2016 年美国图书馆状况报告》中关于高校图书馆的现状描述相呼应。

结果表明，较人工调研方法，基于文本数据挖掘模型的统计方法可以更直观、快速地给出各个图书馆网站导航的共性和特异性的直观分析结果，可供国内建设一流大学图书馆网站参考。

参考文献：

[1] 相丽玲，董小燕，屈宝强. 从网页设计看高校图书馆的网站建设[J]. 情报学报，2004，23(2)：204-208. (Xiang Liling, Dong Xiaoyan, Qu Baoqiang. Study of the Website Construction of Colleges and Universities Library from the Web Page Design[J]. Journal of the China Society for Scientific and Technical Information, 2004, 23(2): 204-208.)

[2] 范爱红, 邵敏, 赵阳. 大学图书馆网站设计理念的探析与实践——清华大学图书馆网站改版案例研究[J]. 大学图书馆学报, 2006, 24(3): 38-42. (Fan Aihong, Shao Min, Zhao Yang. Discussion and Practice on the Principles of University Library Website Design——A Case Study of the Tsinghua University Library Website Redesign[J]. Journal of Academic Libraries, 2006, 24(3): 38-42.)

[3] 范爱红, 姚飞, 姜爱蓉. 清华大学图书馆新版网站的设计特色与读者调查分析[J]. 大学图书馆学报, 2011, 29(5): 66-69. (Fan Aihong, Yao Fei, Jiang Airong. The Features of the New Website of Tsinghua University Library and the Website Survey Analyses[J]. Journal of Academic Libraries, 2011, 29(5): 66-69.)

[4] 叶鹰. 美国一流大学及其图书馆调研报告[J]. 大学图书馆学报, 2002, 20(3): 5-8. (Ye Ying. A Survey on the Top Universities and Their Libraries in the United States[J]. Journal of Academic Libraries, 2002, 20(3): 5-8.)

[5] 谢丽娟, 郑春厚. 美国高校图书馆社会服务发展现状及启示[J]. 中国图书馆学报, 2009, 35(2): 93-97. (Xie Lijuan, Zheng Chunhou. Social Services of Academic Libraries in USA: Reality and Inspirations[J]. Journal of Library Science in China, 2009, 35(2): 93-97.)

[6] Semantic Studios. User Experience Design [EB/OL]. [2016-10-15]. http://semanticstudios.com/user_experience_design/.

[7] Vector Space Model [EB/OL]. [2016-10-15]. https://en.wikipedia.org/wiki/Vector_space_model.

[8] American Library Association. The State of American's Libraries [R/OL]. [2016-10-16]. http://www.ala.org/news/sites/ala.org.news/files/content/state-of-americas-libraries-2016-final.pdf.

[9] TAGUL [CP/OL]. [2016-10-15]. https://tagul.com/.

作者贡献声明:

尹相权: 提出研究思路, 设计研究方案, 进行实验, 采集、清洗和分析数据, 起草论文;
李书宁: 设计研究方案, 论文修改。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: yinxq@lib.bnu.edu.cn。

[1] 尹相权, 李书宁. 美国一流大学图书馆网站导航文本.xlsx. 美国一流大学图书馆网站导航文本及 VSM 结果.

[2] 尹相权, 李书宁. WebsiteAnalysis.rar. VSM 源码 for 网站导航数据.

收稿日期: 2016-11-22
收修改稿日期: 2017-01-24

Analyzing Website Navigation Features of Top U.S. Academic Libraries

Yin Xiangquan Li Shuning
(Beijing Normal University Library, Beijing 100875, China)

Abstract: [**Objective**] This paper studies the navigation features of top academic library websites from the United States, aiming to improve the services of their Chinese counterparts. [**Methods**] First, we identified library websites of the top 15 U.S. universities and downloaded their navigation texts. Second, we analyzed the similarities and differences among these texts with tag cloud and Vector Space Model. Finally, we examined our findings with the "2016 State of America's Libraries Report". [**Results**] The proposed method was intuitive and generated analysis results fast, which could be further processed with text mining techniques. [**Limitations**] Only retrieved the first and, second levels of navigation as well as titles of the homepages. [**Conclusions**] The proposed model provides useful information for the academic libraries in China.

Keywords: U.S. Academic Libraries Website Navigations VSM Investigation and Analysis